テキストマイニングによる文書分析

―内閣総理大臣の所信表明演説を元に―

清水勇吉[i]

1. はじめに

テキストマイニングとは、「大量のデータから解析によって有用な情報を抽出する“データマイニング（data mining）”の一種」と田野村（2011）で言われるように、データマイニングの下位区分であり、対象をテキストに限定したものである。細かな定義に関しては研究者によって異なるが、本稿では簡略的に「種々の文書データを計量的に加工・分析するための一手法」と定義する。文学作品のような文章の集合のほか、テキスト化されたものであればアンケートなどによって得られた自由回答形式のもの、談話資料もその対象とする。

テキストマイニングの目的は、テキストを形態素に分解、集計し、統計処理をすることでそのテキストに潜む傾向を見出すことである。主観によりがちな文書データの分析も、計量化によって客観的視点を取り入れることができる。

本稿は内閣総理大臣の就任時所信表明演説を例に、四名の総理大臣の就任したそれぞれの時期の社会情勢を、テキストマイニングを通して見出そうとするものである。

2. テキストマイニングを行う

テキストマイニングを行うためには、専用のソフトが必要となる。テキストの「形態素解析」、それによって得られた数量化されたデータに対する「統計処理」、そしてそこから分析や考察を行うことまでがテキストマイニングの一連の流れであるが、そこに不可欠なソフトを簡単に以下に挙げる。

2. 1 解析器

英語などのように分かち書きでない日本語のテキストを扱う場合には、形態素に分解することが必須であり、それゆえにテキストマイニングソフトにはどれも解析器が積まれている。フリーで公開されている形態素解析器には

JUMAN[ii]、ChaSen[iii]（茶筌）、MeCab[iv]（和布蕪）などが、構文解析器には CaboCha[v] や KNP[vi] などがある。これらはインターネットでダウンロードするなどして利用が可能だが、有料のテキストマイニングソフトの解析器はどのようなものを用いているかは公開されていないことが多い。

2. 2 テキストマイニングソフト

テキストマイニングソフトには上に挙げたような解析器を利用することで形態素解析を行う。しかし形態素に分解しただけでは不十分で、そのままでは単なる自然言語処理に留まる。そこから統計処理を行うことではじめてテキストマイニングと言える。統計処理に関してはソフトウェア自身の機能で、また他の統計ソフトと連携することで成る。

インターネット上で公開されているテキストマイニングソフトには TTM[vii] や KH Coder[viii] などがあり、これらは MeCab や ChaSen などの形態素解析器、CaboCha などの構文解析器と連携して作動する。統計処理として TTM は語やタグ、またテキストの各組み合わせでクロス表を作成する、KH Coder は語の集計から散布図の作成、また対応分析なども行うことが、それぞれ可能である。必要に応じてそれぞれのソフトを使い分けることが望ましい。

3. 分析方法

3. 1 使用ソフトウェア

分析に使用するフリーのテキストマイニングソフトは語の集計を主眼に置き TTM を、統計処理に関しては Excel 統計 2008 を用いて対応分析を行うこととする。加えてネットワーク図作成のためにトレンドサーチ 2008[ix]（SSRI 社）も利用する。

3. 2 対応分析

1960 年代にフランスのベンゼクリによって提案されたもので、数理的には数量化Ⅲ類と同様の手法で、またコレスポンデンス分析ともいう。名義変数や順序変数など質的な変数を集計したクロス集計表をもとにして次元縮約を行う手法で、多変量解析の一つである。結果として、類似した項目（表側）同士の距離と、類似した変数（表頭）同士の距離を同時に算出することができる。これら得られた座標データからマップを作成することで視覚化されることが多い。

分析に際して散布図を作成するが、軸それぞれについて語同士の相関関係

を最大にするということで、図の見方としては「似た傾向のものが近くに布置される」ことを念頭に置いて見ることが基本となる。ここで言う「似る」とは“回答の傾向（パターン）”であって“語の意味の傾向”ではない。辞書的意味がよく似た二つの語があったとして、それらの意味がどれだけ似ていようとも回答のされ方がまったく異なるのであれば、離れて布置される。またそのことから、「特徴的なものは他の語から離れて布置される」こと、「全体的に同じような回答傾向にあるもの（回答数の多いもの）は原点付近に布置される」ことが言える。ただしこれらはあくまでも基本的な前提であって、常に適用されるものではないことを注意されたい。

4. 分析

4. 1 分析対象

例として日本の過去四代の内閣総理大臣の就任時の所信演説表明の分析を行う。対象とするのは、

- 第 91 代内閣総理大臣：福田康夫[x]（自由民主党総裁）
- 第 92 代内閣総理大臣：麻生太郎[xi]（自由民主党総裁）
- 第 93 代内閣総理大臣：鳩山由紀夫[xii]（民主党代表）
- 第 94 代内閣総理大臣：菅直人[xiii]（民主党代表）

以上四名の所信演説表明の原文である。首相官邸のホームページ[xiv]に公開されているものを利用した。ゆえに、実際の所信表明時にみられたであろうフィラーや言い淀みなどはここに含まれず、また分析対象ともしないため、ここでは問題としない。

1955 年より続いていた（一部の例外を除く）自由民主党政権が崩れたのは 2009 年であった。就任時の所信表明演説とは前任の総理大臣から引き継いだ政権、世論を受けた上での意思表明であるため、施政方針のみならず政治に対する態度、国民への呼びかけ、各々が重視しているものなど多くの情報を含むものとなる。

この政権交代前後の二名ずつの内閣総理大臣の就任時所信表明演説を扱うことで、政権交代という社会的に大きな事象に係る時期の、日本の社会情勢をみるものとする。

4. 2 テキスト（所信表明演説）の加工と図示

本節では、元になるテキストを数量化し、図表化をおこなう。TTM を用

いて形態素解析と同時にクロス表を作成したものが表1である。ここで同時に単語の表記の統一などを含むデータの整理をおこなっている。

表1　整形後のクロス表

語	福田	菅	鳩山	麻生	計	語	福田	菅	鳩山	麻生	計
国民	16	21	42	22	101	問題	8	3	9	1	21
私	8	28	41	24	101	今	3	5	9	4	21
日本	5	5	37	9	56	不安	4	4	3	9	20
政治	5	14	25	4	48	政策	6	6	6	2	20
実現	9	20	11	5	45	成長	5	6	5	3	19
皆さま	12	10	20	2	44	社会保障	1	16	0	2	19
新しい	2	10	28	2	42	信頼	6	3	7	3	19
必要	7	13	8	5	33	解決	5	3	7	3	18
地域	3	8	15	7	33	支援	2	8	6	1	17
我が国	9	14	2	6	31	地方	10	1	5	1	17
経済	3	14	11	2	30	すべて	6	1	9	1	17
国	4	5	16	4	29	民主党	0	3	1	12	16
改革	8	9	4	6	27	社会	3	4	9	0	16
行政	9	7	7	4	27	内閣	1	11	2	2	16
強い	0	16	5	6	27	将来	8	4	4	0	16
安心	12	5	5	2	24	今後	6	3	6	0	15
人	1	6	15	1	23	取組	5	7	3	0	15
世界	4	3	12	4	23	強化	6	4	4	1	15
課題	2	9	5	5	21	推進	6	6	3	0	15
責任	4	4	5	8	21	雇用	1	8	3	3	15

TTM で形態素解析してクロス表を作成し、整形したあとは Excel 統計 2008 を利用して対応分析を実行する。

対応分析にかける語の数は任意で選択する。ここで重要なのは散布図にすることを念頭に置くことである。図上に表示させる語の数は、ソフトによって上限はあるものの基本的にはいくらでも選択可能ではあるが、散布図にしたときに語が多すぎては図が見づらく、少なすぎても分析に足る結果にはなりがたい。今回はそれらを勘案して、便宜的に 40 語とした。その結果が図1である。第1軸の寄与率は 0.3968、第2軸は 0.3554 であり、累積寄与率は 0.7522 である。第1軸と第2軸それぞれの寄与率の差が大きくないことから、図の見方としてどちらかの軸に拠ることもなく、また累積寄与率が高いことからも図全体を俯瞰するのが適当であると考えられる。

簡単に図1に関して述べておくと、四つの象限に四名がそれぞれ布置されており極端に近くなるようなことがないため、各人の特徴がよく示されている図だといえよう。ただしこれらの位置関係はあくまでも全体からの相対的

な関係によるものであるため、たとえば“福田”と“責任”が離れて布置されているからといって“責任”を軽視しているとは言えず、また“責任”に関して言えば“福田”と“菅”は同じ出現頻度を示しているが（表1）、これも相対的な関係が示されているためである。

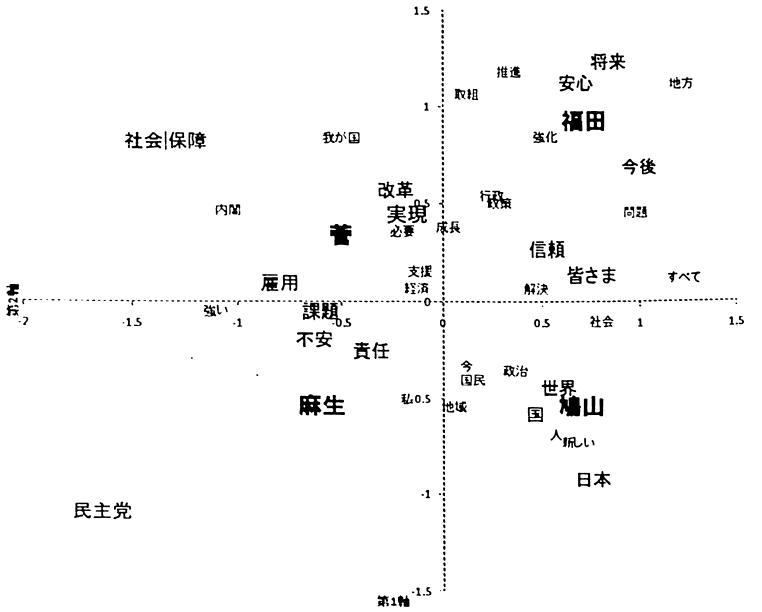

図1　対応分析による散布図

また同時にトレンドサーチ2008というテキストマイニングソフトを利用してネットワーク図を描く。このソフトウェアは有料のものだが、自由回答などの分析にたいへん有用であるため、本発表でも用いることとする。形態素解析など、他のテキストマイニングソフトと基本的な動作は変わらないがビジュアル化におけるレベルは高い。

この図にあらわれる情報は多く、ノード[xv]同士の繋がりや関連度の強さのみならず、ノード間の距離まで計算し布置する。つまり関連度の高いものは近くに、低いものは遠くにそれぞれ布置されるということを意味する。この辺りは対応分析に似たものがある。ただし、正確な距離を図示した場合、ノ

ードの多くは互いに重なりあって見えなくなってしまうため便宜的にずらす事は多い。

図2は、トレンドサーチ2008でキーワードとなった語50語を採用している。対応分析の散布図以上に配置の自由度が高いため、採用語数は少々多くした。

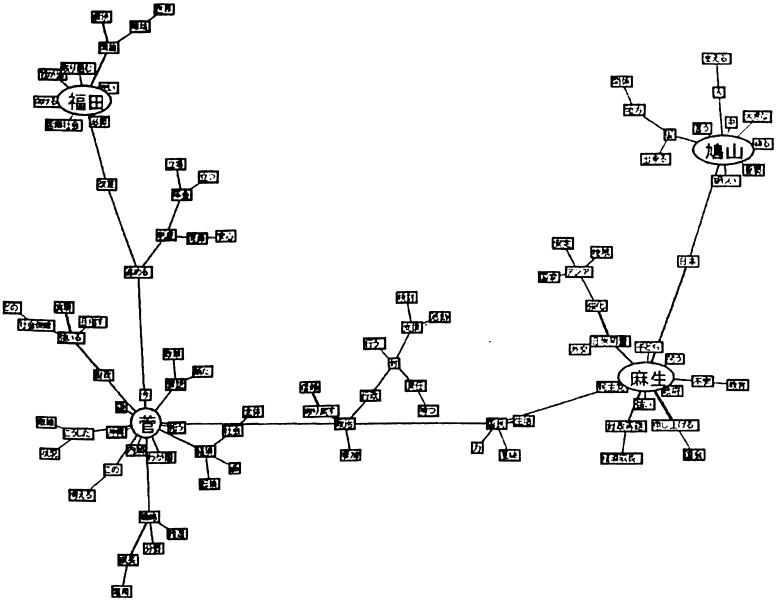

図2　トレンドサーチ2008によるネットワーク図

4.3 分析

以上二つの図から、内閣総理大臣四名の傾向を探る。両図ともに、四名の名前に近い語が相対的ながらそれぞれ重要視されたものと判断できよう。この結果を説明するには、福田氏以前の内閣総理大臣の存在が不可欠である。

まず第87～89代内閣総理大臣として、2001年4月26日から2006年9月26日までの通算1980日もの長期の在任期間を誇った小泉純一郎氏の存在がある。氏に対する政治的社会的評価はさておくとして、第71～73代内閣総理大臣中曽根康弘[xvi]氏以来の1000日を超える在任期間であったことは間違いなく、構造改革や靖国神社への参拝など数々のインパクトを世間に与え

任期満了とともに辞任した。

後継の安倍晋三氏はそれを受けて内閣総理大臣の職に就いている。就任してのち、年金記録問題の浮上、閣僚の諸問題などが重なったことで内閣支持率は下がっていき、また自身が体調を崩すなどしたことからわずか一年の任期で辞任するに至った。在任中の行動、辞任への経緯など、小泉氏と安倍氏の社会的印象に関する明暗がはっきりと分かれる形となった。

以上の流れを経た上で福田康夫氏が次の内閣総理大臣に就任した。小泉氏の長期政権から安倍氏の突然の辞任までの落差を目にしたことによって、国内の、自民党政権に対する不安、不信感が高まっていたことは想像に難くない。だからこそ図1より、福田氏は“安心”や“信頼”といったことばを演説の中で多用することで、国民の現政権に対するネガティブな印象を払拭するとともに信頼を得ようとしたのではないかと推察できる。また図2からは“問題”－“解決”や“取り組む”といった積極的な姿勢も見受けられる。

続いての麻生氏は図1から“課題”や“不安”などが近くに布置されており、安倍氏と同様にわずか一年で任期を終えてしまった福田氏によるものであろう。国民の不安感が反映された可能性も考えられる。図2からは“日米同盟”や“アジア”などがみられることから、自身の外務大臣の経験[vii]からなのか外交に対する関心もうかがえる。しかし麻生氏もマスコミのネガティブ・キャンペーンとも言えるものもあり、世論の批判を受ける形で一年と満たずに政権を交代することとなった。

鳩山氏は自民党との政権交代時により、55年体制以来初の民主党の内閣総理大臣となった。しかし突如野党から与党になったからか、図1をみる限り他の総理大臣と比べて具体的な語が付近に布置されておらず、方針が不明瞭である。表1と併せてみても“国民”や“皆さま”など呼びかけるような語が多いこと程度しか特徴はあらわれない。図2も同様で、あまり特徴的な語が多いとは言いがたい。“地方”との距離から、地方に対する関心があっただろうことは言えようか。鳩山氏は四名の中でも最短の在任期間で菅氏にその座を譲っている。

前総理大臣の菅氏は、前任の鳩山氏と対比される形で図1には“社会保障”や“雇用”など具体的な語が近くに布置されている。“改革”や“実現”なども同様で、前政権で急落した支持率を回復するため、積極的な姿勢を見せている。図2もまた同様に、時勢に合わせたことばを使用している印象がある。ここで興味深いのは、“菅”と“麻生”に関連性があることである。両ノード

間にあるものをみると“信頼”や“取り戻す”があり、前任者の任期が短く、また突然の辞任であったことなどに共通点が存在する。このような面も反映された図になったといえる。

5. おわりに

就任時の所信表明演説は政権の発足時になされるものであり、前任者が世に与えた政権のイメージ、その時代時代の風潮を受けて、内閣総理大臣としての意思、姿勢をあらわすものである。そのテキストのみでも各総理大臣の就任する時期がどういう情勢の中にあったかなど、少なからず特徴を読み取ることができた。

文系の領域では未だ多くない統計分析を加えることで、曖昧な印象に留まることなく客観性の高い分析・考察を行うことができる。本稿では例としてはじめからテキスト化されたものを扱ったが、今後は方言談話資料など、みずからテキスト化する必要のあるもの、また標準的な日本語の辞書にない方言形式の扱いなども含めて、テキストマイニングの適用範囲の拡大を試みたい。

参考文献

石井哲（2002）『テキストマイニング活用法』リックテレコム
上田太一郎（2004（新版第5刷））『新版 Excel でできるデータマイニング入門』同友館
内田治（2010）『数量化理論とテキストマイニング』日科技連出版社
喜田昌樹（2008）『テキストマイニング入門―経営研究での活用法―』白桃書房
鈴木崇史（2008）「総理大臣国会演説における基本的文体特徴量の探索的分析」『計量言語学』26巻4号、pp113-122
林俊克（2002）『Excel で学ぶテキストマイニング入門』オーム社
松村真宏・三浦麻子（2009）『人文・社会科学のためのテキストマイニング』誠信書房
三室克哉・鈴村賢治・神田晴彦（2007）『顧客の声マネジメント―テキストマイニングで本音を「見る」―』オーム社
村上征勝編（2006）『文化情報学入門』勉誠出版

[i] 徳島大学大学院総合科学教育部博士後期課程
[ii] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[iii] http://chasen-legacy.sourceforge.jp/
[iv] http://mecab.sourceforge.net/
[v] http://chasen.org/~taku/software/cabocha/
[vi] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html
[vii] http://mtmr.jp/ttm/
[viii] http://khc.sourceforge.net/
[ix] http://software.ssri.co.jp/fuji/index.html
[x] 在任期間：2007年9月26日-2008年9月24日
[xi] 在任期間：2009年9月16日-2010年6月8日
[xii] 在任期間：2008年9月24日-2009年9月16日
[xiii] 在任期間：2010年6月8日-2011年9月2日
[xiv] http://www.kantei.go.jp/
[xv] 各語や総理大臣の名前を指す。
[xvi] 在任期間：1982年11月27日-1987年11月6日（第一次から第三次まで）
[xvii] 第三次小泉内閣、安倍内閣において外務大臣を歴任している。